



**Birla Institute of Technology & Science, Pilani**  
**Work Integrated Learning Programmes Division**

\*\*\*\*\* \*\*

**Digital Learning Handout**

**Part A: Content Design**

|              |                                                                                                                                   |
|--------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Course Title | Data Management for Machine Learning                                                                                              |
| Course No(s) | AIML* ZG529                                                                                                                       |
| Credit Units | 4                                                                                                                                 |
| Credit Model | 3-1-0<br>(XX Hours of Class-room Instruction + XX Hours of Case-studies/Tutorials/Laboratories + XX Hours of Student Preparation) |
| Instructors  | <b>SUNIL BHUTADA (Lead), PATIL SANDEEP RAMSING</b>                                                                                |
| Version No:  | 1.3                                                                                                                               |
| Date:        | 20 <sup>th</sup> October 2025                                                                                                     |

**Course Description:**

Data Models and Query Languages: Relational, Object-Relational, NoSQL data models; Declarative (SQL) and Imperative (MapReduce) Querying; Data Encoding: Evolution, Formats, Models of dataflow; Machine learning workflow; Data management challenges in ML workflow; Data Pipelines and patterns; Data Pipeline Stages: Data extraction, ingestion, cleaning, wrangling, versioning, transformation, exploration, feature management; Modern Data Infrastructure: Diverse data sources, Cloud data warehouses and lakes, Data Ingestion tools, Data transformation and modelling tools, Workflow orchestration platforms; ML model metadata and Registry, ML Observability, Data privacy and anonymity.

**Course Objectives**

| No         | Course Objective                                                                                                                                                |
|------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>CO1</b> | Introduce students to the data models, storages and querying languages used in data management emphasizing on machine learning aspects                          |
| <b>CO2</b> | Provide required guidance on architecture of modern data platform, usage and types of data pipelines                                                            |
| <b>CO3</b> | Deliver Hands-on exposure to the common techniques, and tools used by data engineers to support build, test, deploy and automate the machine learning pipelines |
| <b>CO4</b> | Expose students to the industry best practices essential to deal with data privacy, metadata and observability                                                  |

**Text Book(s):**

|           |                                                                                          |
|-----------|------------------------------------------------------------------------------------------|
| <b>T1</b> | Fundamentals of Data Engineering: Plan and Build Robust Data Systems by Reis and Housley |
|-----------|------------------------------------------------------------------------------------------|





|           |                                                                                                          |
|-----------|----------------------------------------------------------------------------------------------------------|
| <b>T2</b> | Reliable Machine Learning by Cathy Chen, Niall Richard Murphy, Kranti Parisa, D. Sculley, Todd Underwood |
|-----------|----------------------------------------------------------------------------------------------------------|

**Reference Book(s) & other resources:**

|           |                                                           |
|-----------|-----------------------------------------------------------|
| <b>R1</b> | Designing Data-Intensive Applications by Martin Kleppmann |
| <b>R2</b> | Data Pipelines Pocket Reference by Densore                |
| <b>R3</b> | Building Machine Learning Pipelines by Hapke, Nelson      |

**Learning Outcomes: Students will be able to**

|            |                                                                                                                                                                                 |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>LO1</b> | Explain and differentiate the necessity, position and role of data management components appearing in the modern data stacks                                                    |
| <b>LO2</b> | Analyse and evaluate data ingestion, flow, and processing challenges and propose suitable architectural solutions.                                                              |
| <b>LO3</b> | Design and implement dataflows for machine learning pipelines using contemporary data engineering tools                                                                         |
| <b>LO4</b> | Apply the acquired conceptual data management knowledge and practices over a real-world machine learning workflow addressing the model metadata, privacy and monitoring aspects |

**Modular Content Structure**

**M1 Foundations of data management**

- Data Representations
- Data Management Fundamentals
- Data Architectures

**M2 Modern Data Platform**

- Data Pipelines
- Modern Data Infrastructure and DataOps
- ML Workflow/lifecycle

**M3 Data Management in ML Workflow**

- Data Ingestion
- Data Profiling and Validation
- Analytics Engineering
- Workflow Orchestration and Automation
- Feature Preparations
- ML Experimentation and Metadata

**M4 Frontier Practices in Data & ML Systems**

- Distributed and Scalable Data Processing
- Data Pipelines for LLMs
- Data Privacy, Governance and Responsible ML
- Data and ML Observability





### **Part B: Learning Plan**

| <b>Contact Session</b> | <b>List of Topic Title</b>          | <b>Sub-Topics</b>                                                                                                                                                                                                                                                                          | <b>Reference</b>                 |
|------------------------|-------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|
| 1                      | <b>Data Representations</b>         | <ul style="list-style-type: none"><li>• Data Formats</li><li>• Data Models and Query Languages</li><li>• Data Encoding / Serializations</li><li>• Data Processing</li></ul>                                                                                                                | R1 Chapter 2,3,4                 |
| 2                      | <b>Data Management Fundamentals</b> | <ul style="list-style-type: none"><li>• Data Management Principles</li><li>• Data Management Components</li><li>• Data Management Roles</li><li>• Phases of Data</li></ul>                                                                                                                 | T2 chapter 2                     |
| 3                      | <b>Data Architectures</b>           | <ul style="list-style-type: none"><li>• Data Architectures - Defined</li><li>• Types of Data Architecture: Traditional vs Modern</li><li>• Emerging Data Architectures</li><li>• Lakehouse vs Data Warehouse vs Data Lake</li><li>• Generalized Architecture of Big Data Systems</li></ul> | T1 Chapter 3,4<br>R2 Chapter 2,3 |
| 4                      | <b>Data Pipelines</b>               | <ul style="list-style-type: none"><li>• Modes of Data Flow</li><li>• Modern Data Stack</li><li>• Data Pipelines</li><li>• ETL, ELT</li></ul>                                                                                                                                               | T1 Chapter 6                     |





|   |                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                              |
|---|-----------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------|
|   |                                               | <ul style="list-style-type: none"><li>• Unified data flow</li><li>• Tools / Platforms</li></ul>                                                                                                                                                                                                                                                                                                                                                                          |                              |
| 5 | <b>Modern Data Infrastructure and DataOps</b> | <ul style="list-style-type: none"><li>• Modern Data Infrastructure</li><li>• Data Science Infrastructure</li><li>• Data Serving for Analytics and ML</li><li>• Automation in Data Infrastructure Management</li><li>• DevOps Foundation</li><li>• Continuous Training (CT) and Continuous Deployment (CD) pipelines</li><li>• Intro to DataOps and Observability</li><li>• Case Study: Modern ML Data Infrastructure (integrating DataOps &amp; Observability)</li></ul> | T1 Chapter 6                 |
| 6 | <b>ML lifecycle and Workflow</b>              | <ul style="list-style-type: none"><li>• ML Lifecycle</li><li>• Three Levels of ML Software</li><li>• Data: Data Engineering Pipelines</li><li>• Model: Machine Learning Pipelines</li><li>• Different forms of ML workflows</li></ul>                                                                                                                                                                                                                                    | T2 Chapter 1<br>R3 Chapter 1 |





|   |                                      |                                                                                                                                                                                                                                                                                                                                                                    |                 |
|---|--------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
|   |                                      | <ul style="list-style-type: none"><li>• Code: Deployment Pipelines</li><li>• DataOps → MLOps relationship</li><li>• Case survey: model deployment pipeline failures</li></ul>                                                                                                                                                                                      |                 |
| 7 | <b>Data Collection and Ingestion</b> | <ul style="list-style-type: none"><li>• Data Sources</li><li>• Data Generation in Source Systems</li><li>• Data Ingestion</li><li>• Data contracts, Lineage and Governance</li><li>• Data Ingestion Considerations</li><li>• Ways to Ingest Data</li></ul>                                                                                                         | T1 Chapter 5, 7 |
| 8 | <b>Data Profiling and Validation</b> | <ul style="list-style-type: none"><li>• Common Problems with Data</li><li>• Data Quality Dimensions</li><li>• Bias and Fairness</li><li>• Data Leakage</li><li>• Data Skew and Drift</li><li>• Drift detection metrics</li><li>• Profiling for ML readiness</li><li>• Filter / Selection / Sampling, Validation</li><li>• Data Validation Approaches and</li></ul> | R3 Chapter 4    |





|    |                                              |                                                                                                                                                                                                                                                                                                                                                                                      |                                                 |
|----|----------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------|
|    |                                              | <p>Frameworks</p> <ul style="list-style-type: none"> <li>Automated Data Quality Testing in Pipelines (Great Expectations, Soda Core, Deequ)</li> <li>pre-ingestion validation vs pre-training validation</li> </ul>                                                                                                                                                                  |                                                 |
| 9  | <b>Analytics Engineering</b>                 | <ul style="list-style-type: none"> <li>Data Integration</li> <li>Transformation and Modeling Patterns (ELT, dbt-style)</li> <li>Testing in Data Transformations</li> <li>Data Transformation</li> <li>Data Transformation with dbt</li> <li>Data Versioning</li> <li>Data Lineage and Reproducibility</li> <li>Documentation of Data Models (dbt docs, DataHub, Amundsen)</li> </ul> | <p>Classroom discussions<br/>Demonstrations</p> |
| 10 | <b>Workflow Orchestration and Automation</b> | <ul style="list-style-type: none"> <li>Need for Orchestration</li> <li>Data Pipeline vs Workflow</li> <li>Core Concepts</li> <li>Common Orchestration Tools - Airflow, Prefect, Dagster overview.</li> <li>Designing an Orchestrated Workflow</li> </ul>                                                                                                                             | <p>Classroom discussions<br/>Demonstrations</p> |





|    |                                          |                                                                                                                                                                                                                                                                                                                                                                                                        |                       |
|----|------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
|    |                                          | <ul style="list-style-type: none"> <li>Integrating DataOps with Orchestration</li> <li>Best Practices</li> </ul>                                                                                                                                                                                                                                                                                       |                       |
| 11 | <b>Feature Preparations</b>              | <ul style="list-style-type: none"> <li>Feature life cycle</li> <li>Data Annotation / labelling</li> <li>Data augmentation and Data Synthesis</li> <li>Common Feature Engineering Operations</li> <li>Feature Importance</li> <li>Feature Generalization</li> <li>Feature Reuse and Sharing Practices</li> <li>Feature Stores</li> <li>Feature Governance</li> <li>Engineering Good Features</li> </ul> | T2 Chapter 4          |
| 12 | <b>ML Experimentation &amp; Metadata</b> | <ul style="list-style-type: none"> <li>Model training &amp; experimentation</li> <li>Model Analysis &amp; Validation</li> <li>ML Metadata Store</li> <li>Dataset, Feature, Label, Pipeline metadata</li> <li>ML Experiment Tracking</li> <li>Experiment tracking frameworks (MLflow, Neptune, Weights &amp;</li> </ul>                                                                                 | Classroom discussions |





|    |                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                    |
|----|---------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|
|    |                                                   | <p>Biases)</p> <ul style="list-style-type: none"><li>• Model Registry and Versioning</li><li>• Metadata governance and lineage tracking</li></ul>                                                                                                                                                                                                                                                                                              |                                                    |
| 13 | <b>Distributed &amp; Scalable Data Processing</b> | <ul style="list-style-type: none"><li>• Big Data Analytics landscape</li><li>• Distributed and Parallel data processing principles</li><li>• In-memory and stream processing paradigms</li><li>• Modern distributed frameworks - Spark Structured Streaming, Flink, Ray, Dask</li><li>• Cloud-native data platforms (Snowflake, BigQuery, Databricks)</li><li>• Batch vs. Streaming trade-offs for ML training vs inference pipeline</li></ul> | <p>Classroom discussions</p> <p>Demonstrations</p> |
| 14 | <b>Data Pipelines for LLMs</b>                    | <ul style="list-style-type: none"><li>• Why LLM Data Pipelines are Different?</li><li>• Types of LLM Pipelines</li><li>• Core Components of an LLM Data Pipeline</li><li>• Data Pipeline Architecture for LLM Systems</li><li>• Key Engineering Considerations</li><li>• RAG pipeline frameworks:</li></ul>                                                                                                                                    | <p>Classroom discussions</p> <p>Demonstrations</p> |





|    |                                                      |                                                                                                                                                                                                                                                                                                                                                                                       |                               |
|----|------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------|
|    |                                                      | LangChain, LlamaIndex, Haystack                                                                                                                                                                                                                                                                                                                                                       |                               |
| 15 | <b>Data Privacy, Governance &amp; Responsible ML</b> | <ul style="list-style-type: none"> <li>• Data Privacy Landscape and Regulations</li> <li>• Governance frameworks: DPDP Act (India), GDPR, CCPA — global compliance perspective</li> <li>• Privacy-Preserving ML Paradigms (Differential Privacy, Federated Learning)</li> <li>• Anonymization and Synthetic Data</li> <li>• Encrypted ML and Secure Computation</li> </ul>            | T2 Chapter 6<br>R3 Chapter 14 |
| 16 | <b>Data and ML Observability</b>                     | <ul style="list-style-type: none"> <li>• Why Observability Matters?</li> <li>• Five Pillars of Data Observability</li> <li>• Practical Implementation</li> <li>• Connecting to DataOps</li> <li>• From Data to ML Observability</li> <li>• Common ML System Failures</li> <li>• Monitoring Metrics</li> <li>• Tooling Landscape</li> <li>• Unified Observability Framework</li> </ul> | T2 Chapter 9                  |



## **Experiential Learning Components:**

Describe objective, outcome of Experiential Learning Component and the lab infrastructure needed (virtual, remote, open source etc..) number of lab exercises needed, etc.

1. Lab work: 04
2. Project work: 01
3. Case Study: 02
4. Simulation: 0
5. Work Integrated Learning Assignment: 0
6. Design work/ Field work: 0

## **Objective of Experiential Learning Component:**

- Design and implement data flows using databases, REST/RPC, and message brokers
- Build a modern data stack with ELT pipelines, cloud storage, transformation tools, and BI platforms
- Manage machine learning model metadata using MLFlow or Neptune for tracking and reproducibility
- Construct a machine learning pipeline integrating data versioning, feature stores, and prediction services

## **Scope of Experiential Learning Component:**

- Identification and usage of various tools / frameworks available for building an end-to-end data plus machine learning pipeline

## **Lab Infrastructure:**

- CSIS Remote Labs / Local setups

## **List of Experiments:**

| Exp No. | Experiment Title                                                                                                                                                                                                                                                             | Reference to handout module/session |
|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------|
| 1.      | Design and implement the simple data flows involving various data formats<br><br>Modes of data flows <ul style="list-style-type: none"> <li>• Through Databases - use SQL / Custom Program to read/write into databases</li> <li>• Through REST/RPC - Synchronous</li> </ul> | M1                                  |





|    |                                                                                                                                                                                                                                                                                                                     |    |
|----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
|    | <p>mechanism for data exchange</p> <ul style="list-style-type: none"> <li>Through Message Brokers / Queues – Asynchronous mechanism for data exchange</li> </ul>                                                                                                                                                    |    |
| 2. | <p>Build a Modern Data Stack</p> <p>Components</p> <ul style="list-style-type: none"> <li>a fully managed ELT data pipeline</li> <li>a cloud-based columnar warehouse or data lake as a destination</li> <li>a data transformation tool</li> <li>A business intelligence or data visualization platform.</li> </ul> | M2 |
| 3. | <p>Manage Machine Learning Model Metadata using MLFlow / Neptune</p> <p>Components</p> <ul style="list-style-type: none"> <li>Projects</li> <li>Experiments</li> <li>Model metadata</li> <li>Model tracking / logging</li> <li>Model Registry</li> </ul>                                                            | M3 |
| 4. | <p>Construct a Machine Learning Pipeline with Data Versioning Tool</p> <p>Components</p> <ul style="list-style-type: none"> <li>Data Pipeline</li> <li>Data Versioning Tool</li> <li>Feature Store</li> <li>ML Pipeline</li> <li>Prediction Service</li> </ul>                                                      | M3 |





## **Evaluation Scheme:**

**Legend:** EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

| Evaluation Component | Name<br>(Quiz, Lab, Project, Mid-term exam, End semester exam, etc.) | Type<br>(Open book, Closed book, Online, etc.) | Weight | Duration | Day, Date, Session, Time |
|----------------------|----------------------------------------------------------------------|------------------------------------------------|--------|----------|--------------------------|
| EC - 1               | Assignment 1                                                         | Take Home / Online                             | 20%    | 21 days  | August 10-20, 2026       |
|                      | Quiz 1                                                               |                                                | 10%    | 7 days   |                          |
| EC - 2               | Mid-Semester Test                                                    | Closed Book                                    | 30%    | 2 hours  | 20/09/2026 (FN)          |
| EC - 3               | Comprehensive Exam                                                   | Open Book                                      | 40%    | 2½ Hours | 06/12/2026 (FN)          |

EC1\* (20% - 30%): Quiz (optional): 5-10 %, Lab Assignment/Assignment: 20% - 30%

Syllabus for Mid-Semester Test (Closed Book): Topics in Contact session: 1 to 8

Syllabus for Comprehensive Exam (Open Book): All topics

## **Important Links and Information:**

**eLearn Portal:** <https://elearn.bits-pilani.ac.in>

Students must visit the eLearn portal regularly and stay updated with the latest announcements and deadlines.

**Contact Sessions:** Students should attend the online lectures as per the schedule provided on the eLearn portal.

## **Evaluation Guidelines:**

1. EC-1 consists of either two Assignments or three Quizzes. Students will attempt them through the course pages on the eLearn portal. Announcements will be made on the portal in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: "open book" means text/ reference books (publisher copy only) and does not include any other learning material. No other learning material will be permitted during the open book examinations. For Detailed Guidelines refer to the attached document.

[EC3 Guidelines](#)





**BITS Pilani**

Pilani | Dubai | Goa | Hyderabad

4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam, which will be made available on the eLearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self-study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignments/Quizzes, Mid-Semester Tests and Comprehensive Exams according to the evaluation scheme provided in the handout.

\*\*\*\*\*

